

USO DE K-MEANS COMO HERRAMIENTA DE IA PARA LA PREVENCIÓN DE RIESGOS Y SÍNTOMAS RESPIRATORIOS EN POBLACIONES TRABAJADORAS VULNERABLES COLOMBIANAS

USE OF K-MEANS AS AN AI TOOL FOR THE PREVENTION OF RESPIRATORY RISKS AND SYMPTOMS IN VULNERABLE WORKING POPULATIONS IN COLOMBIA

Yenny Andrea ROZO SILVA¹, Leidy Isabel CALDERÓN SIERRA², Ana DELGADO-GARCÍA³,
Juan José AGÚN-GONZÁLEZ⁴ y Raúl AGUILAR ELENA⁵

¹ *Facultad de Ciencias Empresariales. Corporación Universitaria Minuto de Dios, Grupo de Investigación Ambientes Laborales Seguros, Saludables y Sostenibles (GIALSSS). Bogotá, Colombia*

✉ yrozo@uniminuto.edu

🌐 <https://orcid.org/0000-0002-3876-2372>

² *Facultad de Ciencias Empresariales. Corporación Universitaria Minuto de Dios, Grupo de Investigación Ambientes Laborales Seguros, Saludables y Sostenibles (GIALSSS). Bogotá, Colombia*

✉ icalderon@uniminuto.edu

🌐 <https://orcid.org/0000-0002-1056-901X>

³ *Grupo de Investigación BISITE, Universidad de Salamanca. Salamanca, España*

✉ a.delgado@usal.es

🌐 <https://orcid.org/0009-0003-7079-5507>

⁴ *Grupo de Investigación Prevención de Riesgos Laborales y Salud Laboral (GPRL), Cátedra PREVION. Universidad Internacional de Valencia (VIU). Valencia, España*

✉ jjagun@universidadviu.com

🌐 <https://orcid.org/0000-0002-2677-7198>

⁵ *Grupo de Investigación Prevención de Riesgos Laborales y Salud Laboral (GPRL), Cátedra PREVION. Universidad Internacional de Valencia (VIU). Valencia, España*

✉ raul.aguilar.e@professor.universidadviu.com

🌐 <https://orcid.org/0000-0001-8656-8155>

RESUMEN: El reciclaje informal en Colombia es una actividad ambientalmente necesaria, desarrollada en condiciones de elevada vulnerabilidad social y sanitaria, con exposición crónica

a riesgos biológicos y particulados y escaso acceso a vigilancia en salud laboral. En este contexto, la inteligencia artificial (IA) ofrece una oportunidad para identificar perfiles de riesgo ocultos en poblaciones sin datos etiquetados ni diagnósticos previos. Se realizó un estudio transversal en 179 recicladores informales de Bogotá, Tocancipá, Cajicá y Funza. Los síntomas respiratorios se evaluaron mediante el cuestionario ATS-DLD-78-A1 y la exposición y percepción del riesgo biológico mediante el cuestionario ERBIO. Los datos se procesaron en R y se analizaron mediante seis técnicas de *clustering* no supervisado (K-means, K-modes, jerárquico, DBSCAN, modelos de mezcla gaussiana y Fuzzy C-means), comparadas mediante métricas de validación interna. El algoritmo K-means mostró el mejor rendimiento global (índice de silueta = 0,639) e identificó tres clústeres: un grupo envejecido y altamente vulnerable con elevada prevalencia de síntomas respiratorios; un grupo joven con baja sintomatología; y un grupo intermedio, predominantemente femenino, con riesgo moderado. Otros métodos detectaron subgrupos extremos, pero con menor estabilidad e interpretabilidad. El *clustering* no supervisado, especialmente K-means, permite identificar perfiles diferenciados de riesgo respiratorio y orientar estrategias preventivas específicas. Integrada en la salud pública, la IA contribuye a la detección temprana del riesgo, la reducción de desigualdades y el avance hacia una prevención de precisión en salud laboral.

Palabras clave: inteligencia artificial; clustering no supervisado; síntomas respiratorios; trabajadores informales.

ABSTRACT: Informal recycling in Colombia is an environmentally necessary activity carried out under conditions of high social and health vulnerability, with chronic exposure to biological and particulate risks and limited access to occupational health surveillance. In this context, artificial intelligence (AI) offers an opportunity to identify hidden risk profiles in populations without labeled data or prior diagnoses. A cross-sectional study was conducted among 179 informal recyclers from Bogotá, Tocancipá, Cajicá, and Funza. Respiratory symptoms were assessed using the ATS-DLD-78-A1 questionnaire, and biological exposure and risk perception were evaluated using the ERBIO questionnaire. Data were processed in R and analyzed using six unsupervised clustering techniques (K-means, K-modes, hierarchical clustering, DBSCAN, Gaussian mixture models, and Fuzzy C-means), which were compared using internal validation metrics. The K-means algorithm showed the best overall performance (silhouette index = 0.639) and identified three clusters: an older, highly vulnerable group with a high prevalence of respiratory symptoms; a younger group with low symptomatology; and an intermediate, predominantly female group with moderate risk. Other methods detected extreme subgroups but demonstrated lower stability and interpretability. Unsupervised clustering, particularly K-means, enables the identification of differentiated respiratory risk profiles and the design of targeted preventive strategies. When integrated into public health, AI contributes to early risk detection, the reduction of inequalities, and progress toward precision prevention in occupational health.

Keywords: artificial intelligence; unsupervised clustering; respiratory symptoms; informal workers.

1. CONTEXTO Y JUSTIFICACIÓN

El reciclaje informal en Colombia constituye una actividad ambientalmente necesaria, pero desarrollada en condiciones de elevada vulnerabilidad social y sanitaria. Miles de personas dependen de esta labor sin protección social, expuestas de forma continuada a contaminantes y sin acceso efectivo a sistemas de vigilancia en salud [1,3]. Se trata de una población heterogénea y envejecida que refleja un patrón estructural de exclusión laboral. En este contexto, la inteligencia artificial ofrece un enfoque innovador para identificar perfiles de riesgo y patrones de enfermedad. El uso de algoritmos no supervisados, como K-means, permite analizar poblaciones diversas sin imponer categorías previas, aportando evidencia útil para la orientación de políticas públicas en salud laboral.

2. MATERIALES Y MÉTODOS

El estudio incluyó 179 recicladores informales de Bogotá, Tocancipá, Cajicá y Funza, seleccionados mediante muestreo no probabilístico por conveniencia. Se aplicaron el cuestionario ATS-DLD-78-A1 para síntomas respiratorios y el cuestionario ERBIO para exposición y percepción de riesgo biológico, ambos con adecuada consistencia interna ($\alpha > 0,70$). Las variables sociodemográficas y de salud se procesaron en R, descartando varianza nula y recodificando variables categóricas. Posteriormente se aplicaron seis técnicas de clustering (K-means, K-modes, jerárquico, DBSCAN, GMM y Fuzzy C-means), comparadas mediante métricas de validación interna (silueta, conectividad y Dunn).

3. RESULTADOS PRINCIPALES

El algoritmo K-means mostró el mejor rendimiento global (índice de silueta = 0,639), evidenciando una adecuada separación entre grupos. El número óptimo de clústeres fue $k = 3$, lo que permitió identificar tres perfiles poblacionales diferenciados. El Clúster 1 (23,8 %) correspondió a personas de mayor edad (media 64,8 años), con alta prevalencia de síntomas respiratorios y afiliación sanitaria subsidiada, reflejando un perfil de elevada vulnerabilidad. El Clúster 2 (17,9 %) agrupó a personas jóvenes (media 26,8 años), mayoritariamente hombres, con baja escolaridad y escasa sintomatología, representando un grupo laboralmente precario pero funcional. El Clúster 3 (20,4 %) incluyó adultos intermedios (media 46 años), con predominio femenino y presencia moderada de síntomas, asociado a una transición laboral marcada por riesgos de género. Otros métodos (K-modes, DBSCAN) identificaron subgrupos extremos, pero con menor estabilidad y consistencia interna. En conjunto, K-means proporcionó una segmentación clara, equilibrada y replicable, facilitando la interpretación preventiva. Los patrones sintomáticos por clúster se representaron mediante gráficos radar comparativos.

Tabla 1. Métricas de validación interna de los algoritmos de clustering aplicados

Method	Silhouette	Calinski–Harabasz	Davies–Bouldin	Dunn	Noise Ratio	Interpretation
K-Means	0.64	771.18	0.55	0.04	—	Compact, well-separated clusters
Hierarchical	0.64	768.34	0.55	0.09	—	Similar robust structure
GMM	0.29	12.94	3.48	0.15	—	Moderate overlap
K-Modes	—	29.22	2.32	0.03	—	Weak categorical structure
DBSCAN	0.08	10.41	3.94	0.01	0.06	Irregular, noise-tolerant
Fuzzy C-Means	0.04	7.08	5.38	0.01	—	Very diffuse boundaries

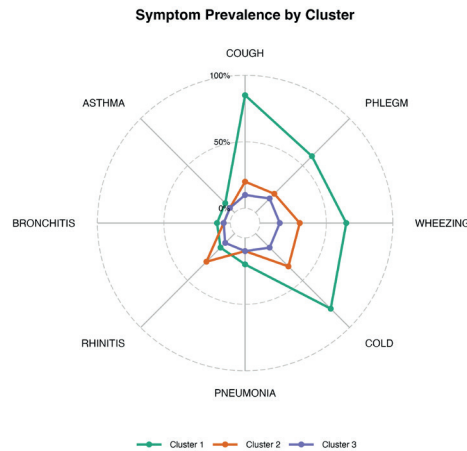


Figura 1. Distribución de síntomas respiratorios por clúster identificado mediante K-means.

4. DISCUSIÓN

Los resultados confirman el potencial del *clustering* como técnica de inteligencia artificial aplicada a la salud laboral. En particular, K-means permitió identificar perfiles de riesgo respiratorio no detectables mediante enfoques estadísticos tradicionales, resultando especialmente adecuado en contextos sin datos etiquetados ni diagnósticos previos, como las poblaciones trabajadoras informales.

La Figura 2 muestra un patrón consistente entre algoritmos: la presencia de al menos un clúster con alta prevalencia de síntomas respiratorios clave frente a otros grupos más

homogéneos y de menor riesgo [2,4]. Aunque métodos como GMM o Fuzzy C-means ofrecen transiciones más graduales y DBSCAN identifica casos extremos, K-means proporciona una segmentación clara, estable e interpretable.

Desde una perspectiva aplicada, K-means destaca por su equilibrio entre simplicidad computacional, robustez y facilidad de transferencia del conocimiento, facilitando el diseño de intervenciones preventivas dirigidas. Más allá del plano técnico, estos resultados evidencian el valor de la IA para visibilizar vulnerabilidades en poblaciones envejecidas, aportando una herramienta analítica con implicaciones éticas y de equidad en salud laboral y salud pública.

5. CONCLUSIONES

Este estudio confirma que la aplicación de inteligencia artificial mediante técnicas de *clustering* no supervisado permite identificar perfiles diferenciados de riesgo respiratorio en poblaciones trabajadoras informales, tradicionalmente invisibilizadas por los sistemas de vigilancia en salud. La segmentación obtenida facilita el diseño de intervenciones preventivas ajustadas a las características demográficas, laborales y sanitarias de cada grupo, superando el enfoque uniforme tradicional. El clúster más vulnerable requiere vigilancia respiratoria intensiva y medidas inmediatas de reducción de la exposición; el clúster de transición demanda acciones específicas con perspectiva de género; y el clúster resiliente se beneficia de estrategias de prevención temprana basadas en formación y control del riesgo. Integrada en políticas públicas y estrategias comunitarias, la inteligencia artificial se consolida como una herramienta eficaz para anticipar el daño, reducir desigualdades estructurales y avanzar hacia un modelo de prevención de precisión en salud laboral y salud pública [1,3].

6. REFERENCES

1. Binion, E., & Gutberlet, J. (2012). The effects of handling solid waste on the wellbeing of informal and organized recyclers: A review of the literature. *International Journal of Occupational and Environmental Health*, 18(1), 43-52. <https://doi.org/10.1179/1077352512Z.0000000001>
2. Ray, M. R., Roychoudhury, S., Mukherjee, G., Roy, S., & Lahiri, T. (2005). Respiratory and general health impairments of workers employed in a municipal solid waste disposal at an open landfill site in Delhi. *International Journal of Hygiene and Environmental Health*, 208(4), 255-262. <https://doi.org/10.1016/j.ijheh.2005.02.001>
3. Black, M., Karki, J., Lee, A. C. K., Makai, P., Baral, Y. R., Kritsotakis, E. I., & Turner, R. (2019). The health risks of informal waste workers in the Kathmandu Valley: A cross-sectional survey. *Public Health*, 166, 10-18. <https://doi.org/10.1016/j.puhe.2018.09.026>
4. Tlotleng, N., Kootbodien, T., Wilson, K., Made, F., Mathee, A., Ntlebi, V., & Naicker, N. (2019). Prevalence of respiratory health symptoms among landfill waste recyclers in the City of Johannesburg, South Africa. *International Journal of Environmental Research and Public Health*, 16(21), Artículo 4277. <https://doi.org/10.3390/ijerph16214277>

